Unsupervised Quality Control of Image Segmentation based on Bayesian Learning

Benoît Audelan, Hervé Delingette

Unsupervised Quality Control of Image Segmentation based on Bayesian Learning

Benoît Audelan and Hervé Delingette

Université Côte d'Azur, Inria, Epione project-team, Sophia Antipolis, France
benoit.audelan@inria.fr

Abstract. Assessing the quality of segmentations on an image database is required as many downstream clinical applications are based on segmentation results. For large databases, this quality assessment becomes tedious for a human expert and therefore some automation of this task is necessary. In this paper, we introduce a novel unsupervised approach to assist the quality control of image segmentations by measuring their adequacy with segmentations produced by a generic probabilistic model. To this end, we introduce a new segmentation model combining intensity and a spatial prior defined through a combination of spatially smooth kernels. The tractability of the approach is obtained by solving a type-II maximum likelihood which directly estimates hyperparameters. Assessing the quality of the segmentation with respect to the probabilistic model allows to detect the most challenging cases inside a dataset. This approach was evaluated on the BRATS 2017 and ACDC datasets showing its relevance for quality control assessment.

Keywords: Quality control · image segmentation · Bayesian learning.

1 Introduction

Quality control of image segmentation is an important task since it impacts the decisions that clinicians or other downstream algorithms can make about the patient. In the case of an automatic pipeline used in a clinical routine, it is therefore of great importance to be able to detect the possible failed segmentations. Many segmentation algorithms follow a supervised learning approach, learning the segmentation task on databases where images and ground truth are jointly available. The main challenges are thus to verify the quality of ground truth segmentations but also to monitor the application of a segmentation algorithm on images for which no ground truth is available. Despite its relevance, the quality control of segmentation has been relatively little studied. In [11], a framework to detect failures in cardiac segmentation based on shape and intensity features has been proposed. A more generic feature based approach has also been explored in [4] where Dice coefficients are predicted by an SVM regressor. Reverse Classification Accuracy (RCA) was proposed in [10], assuming the availability of a subset ground truth dataset. In that case, the proposed segmentation on a new image is compared to the predicted segmentations based on this

subset of reference images, which can result in rejection if discrepancies are too large. This approach was further investigated by [8] on larger databases where they showed the ability to isolate segmentations of poor quality but pointed out the relatively long computation time as a bottleneck. In [7] the authors propose a neural network to directly predict the Dice coefficient. Finally, another deep learning-based approach was introduced in [9] where the uncertainty in the produced segmentation is correlated with its quality.

These methods allow to detect poor segmentations in the absence of ground truth ones but have also some limitations. They are all supervised meaning that they require a subset of segmented data to be considered as "representative ground truth", the size of this subset being potentially large for deep learning-based methods which somewhat defies its purpose. Furthermore, some methods lack interpretability as one may not know why a segmentation has failed.

In this paper, we propose a novel unsupervised approach to automated quality control by comparing segmentations S produced by an algorithm or a human rater to a generic model of segmentation M instead of an arbitrary selected subset of segmentations. This allows to remove the bias related to the subset selection and to monitor the quality of segmentations when few or even no other segmentations are available from a database. In addition, it provides visually interpretable results which could be used for manual corrections of poor cases. To assess the quality of a given segmentation S, we propose to fit a probabilistic generative segmentation model M making simple intensity and smoothness assumptions. The underlying hypothesis is that explainable segmentations correspond to clearly visible boundaries in the image which is well captured by M. On the contrary, segmentations far from M are categorized as challenging as they would require other priors than intensity and smoothness to be explained. These difficult cases can be highlighted by comparing the adequacies between M and S inside a same dataset.

We use a Bayesian framework to estimate automatically all parameters of the segmentation model where the prior probability of a voxel label is defined as a generalized linear model of spatially smooth kernels. Parameter estimation is performed by a sparsity inducing prior for the automatic selection of the number of components of Student mixtures, by solving *type-II maximum likelihood* for controlling the coefficient shrinkage and by performing model selection for the choice of kernels. We show on two public databases, ACDC and BRATS 2017, that our approach is able to monitor the quality of ground truth segmentations but also to indicate the potential performances of segmentations on test data (in the absence of ground truth).

2 Probabilistic Segmentation Framework

Given a segmentation S on an image I, our objective is to produce a smooth contour or surface M close to S which is mostly aligned with visible contours in the image. The estimated segmentation M should not be seen as a surrogate

ground truth, but only as a comparison tool. The adequacy between S and M gives an estimate of the quality of the segmentation S.

We consider a binary image segmentation problem on image I made of N voxels having intensity $I_n \in \mathbb{R}$, $n = 1, \ldots, N$. We introduce for each voxel a binary hidden random variable $Z_n \in \{0, 1\}$ with $Z_n = 1$ if voxel n belongs to the structure of interest.

Appearance models of the foreground and background regions of S are defined respectively by the two image likelihoods $p(I_n|Z_n = 1, \theta_I^1)$ and $p(I_n|Z_n = 0, \theta_I^0)$ where θ_I^0, θ_I^1 are parameters governing those models. In this paper, we consider generic parametric appearance models as variational mixtures of Student-t distributions [1]. The Student-t unlike Gaussian distributions lead to robust mean and covariance estimates and variational Bayesian methods allow to select automatically the number of components. We introduce the appearance probability ratio $r_n(I, \theta_I^0, \theta_I^1) \triangleq p(I_n|Z_n = 1, \theta_I^1) / \left(p(I_n|Z_n = 0, \theta_I^0) + p(I_n|Z_n = 1, \theta_I^1)\right)$ which is the posterior label probability with non-informative prior ($p(Z_n = 1) = 0.5$).

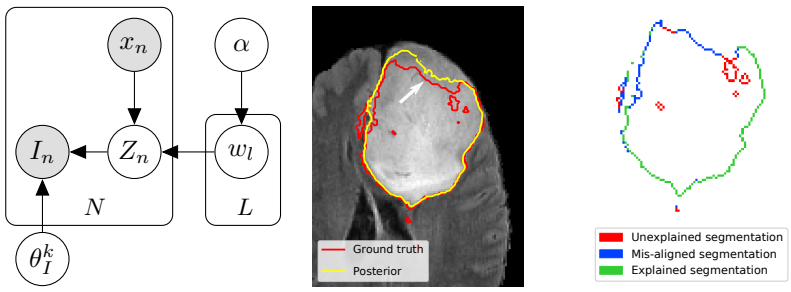
Classical label priors in the literature are based on discrete formulations such as Markov random fields that are relying on labels of neighboring voxels. In this paper, we propose a novel continuous label prior framework defined through a generalized linear model of spatially smooth functions. This approach allows a Bayesian estimation of its parameters $\mathbf{W}$ and produces by construction continuous posterior label distributions. More precisely, the prior probability $p(Z_n = 1)$ is defined as a Bernouilli distribution whose parameter depends on a *spatially random* function $p(Z_n = 1|\mathbf{W}) = \sigma\left(\sum_{l=1}^{L} \Phi_l(\mathbf{x}_n) w_l\right)$ where $\mathbf{x}_n \in \mathbb{R}^d$ is the voxel position in an image of dimension d and $\sigma(u)$ is the sigmoid function $\sigma(u) = 1/\left(1 + \exp(-u)\right)$. The basis $\{\Phi_l(\mathbf{x})\}$ are L functions of space, typically radial basis functions, and $w_l \in \mathbf{W}$ are weights considered as random variables. Thus the prior probabilities of two geometrically close voxels will be related to each other through the smoothness of the function $f(\mathbf{x}_n) = \sum_{l=1}^{L} \Phi_l(\mathbf{x}_n) w_l$.

The smoothness of the label prior $\sigma\left(f(\mathbf{x}_n)\right)$ depends on the choice of the L basis functions $\Phi_l(\mathbf{x}_n)$. The weight vector $\mathbf{W} = (w_1, \ldots, w_L)^T$ is equipped with a zero mean Gaussian prior parameterized by the diagonal precision matrix $\alpha\mathbf{I}$: $p(\mathbf{W}) = \mathcal{N}(0, \alpha^{-1}\mathbf{I})$. Experiments have shown that sharing the same precision α across the weights w_l improves the model stability. Finally, a non-informative prior is chosen for α, $p(\alpha) \propto 1$. The graphical model of the segmentation framework is shown in Fig. 1a.

Once the distribution on $\mathbf{W}$ is known, the prior $p(Z_n = 1)$ can be computed by marginalizing over the weights $\int_{-\infty}^{+\infty} \sigma(\boldsymbol{\Phi}_n \mathbf{W})\, p(\mathbf{W})\, d\mathbf{W}$ writing $\boldsymbol{\Phi}_n = (\Phi_1(\mathbf{x}_n), \ldots, \Phi_L(\mathbf{x}_n))$. It is approximated by $\sum_{l=1}^{L} \Phi_l(\mathbf{x}_n) \boldsymbol{\mu}_l^\star$, where $\boldsymbol{\mu}^\star$ is the mode of $\mathbf{W}$. The posterior label probability $p(Z_n = 1|I, \mathbf{W})$, combining prior and intensity likelihoods, is obtained through Eq. 1:

$$p(Z_n = 1|I, \mathbf{W}) = \frac{r_n(I, \theta_I^0, \theta_I^1) p(Z_n = 1)}{r_n(I, \theta_I^0, \theta_I^1) p(Z_n = 1) + (1 - r_n(I, \theta_I^0, \theta_I^1)) p(Z_n = 0)} \quad (1)$$

Finally, the maximum a posteriori estimate of the segmented structure is obtained as the isosurface $p(Z_n = 1|I, \mathbf{W}) = 0.5$. To estimate prior and hyperprior



(a) (b) (c)

Fig. 1: Graphical model of the framework (1a). Study of the ground truth of the BRATS 2017 challenge with a case of possible under segmentation (1b and 1c).

parameters, we propose to maximize the following log joint probability:

$$\begin{aligned}\log p(I, \mathbf{W}, \alpha, \theta_I) &= \log p(I|\theta_I, \mathbf{W}) p(\mathbf{W}|\alpha) p(\alpha) \\ &= \sum_{n=1}^{N} \log \left(\sum_{k=0}^{1} p(I_n|\theta_I^k, Z_n = k) p(Z_n = k|\mathbf{W}) \right) - \frac{1}{2}\alpha \mathbf{W}^T \mathbf{W} + \frac{L}{2} \log \alpha \end{aligned} \tag{2}$$

3 Bayesian Learning of Prior Parameters

As the final objective is the quality control of the given segmentation S, it is of little interest to work with the whole image and computationally inefficient. We thus restrict the analysis inside a narrow band of width typically between 8 and 30 voxels defined around the boundaries of the foreground region of S.

The method starts with the estimation of the appearance probability ratio r_n for each voxel n. Two variational mixture models of Student-t distributions are fitted, one for the foreground region of S and the other for the background, following the approach of [1]. The sparsity inducing Dirichlet prior over the mixture proportions allows to automatically select the appropriate number of components. Once r_n are known, the problem reduces to estimate the weights $\mathbf{W}$ and the precision α in Eq. 2. Setting $y_n = \boldsymbol{\Phi}_n \mathbf{W}$, the sum in Eq. 2 can be indeed rewritten as $\sum_{n=1}^{N} \log \left[r_n \sigma(y_n) + (1 - r_n)(1 - \sigma(y_n)) \right] + \text{cst}$.

To learn the parameters of the model, we adopt a *type-II maximum likelihood* approach, based on the maximization of the *marginal log likelihood* $\mathcal{L}(\alpha) = \log p(I, \alpha) = \log \int P(I, \mathbf{W}, \alpha) d\mathbf{W}$. The idea is to marginalize out the weight variables such that the maximization is performed only on the precision variable. The marginal log likelihood is intractable but can be approximated through the Laplace approximation of Eq. 2: $\log p(I, \mathbf{W}, \alpha) \approx \log p(I|\boldsymbol{\mu}^\star) + \log p(\boldsymbol{\mu}^\star|\alpha) - \frac{1}{2}(\mathbf{W} - \boldsymbol{\mu}^\star)^T (\boldsymbol{\Sigma}^\star)^{-1} (\mathbf{W} - \boldsymbol{\mu}^\star)$ which corresponds to the following approximation of the posterior probability of the weights: $q(\mathbf{W}) = \mathcal{N}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star) \approx p(\mathbf{W}|I)$.

The computation of the Laplace approximation requires to find the mode $\boldsymbol{\mu}^\star$ of Eq. 2 and to compute the Hessian matrix at the mode. This is done through a Gauss-Newton optimization formulated as an iterative reweighted least squares. This leads to the following expression of the covariance $\boldsymbol{\Sigma}^\star = (\boldsymbol{\Phi}^T \mathbf{B} \boldsymbol{\Phi} + \alpha \mathbf{I})^{-1}$ where $\mathbf{B} = \text{diag}\left(g_1''\left(\sum_i w_i \Phi_i^1\right), \cdots, g_L''\left(\sum_i w_i \Phi_i^L\right)\right)$ is a diagonal matrix, g_n being the function defined as $g_n(x) = \log\left[r_n \sigma(x) + (1 - r_n)\sigma(1 - x)\right]$.

The algorithm alternates between estimating the mean $\boldsymbol{\mu}^\star$ and the covariance $\boldsymbol{\Sigma}^\star$ and updating the precision parameter α through Eq. 3, obtained by taking the derivatives of $\mathcal{L}(\alpha)$ with respect to α, following the approach of [5].

$$\alpha_{new} = \frac{L - \alpha_{old} \,\text{Tr}\left(\boldsymbol{\Sigma}^\star\right)}{\boldsymbol{\mu}^{\star T} \boldsymbol{\mu}^\star} \tag{3}$$

The sketch of the algorithm is provided in Alg. 1.

Algorithm 1: Bayesian Learning algorithm for segmentation

- Define the basis functions and compute their values on the narrow band
- **while** *not converged* **do**
  1) Recompute $\boldsymbol{\Sigma}^\star$ and $\boldsymbol{\mu}^\star$ from the Laplace approximation
  2) Re-estimate α following Eq. 3

end

The choice of the basis functions Φ_l controls the smoothness of the prior. In the remainder, we use a dictionary of Gaussian bases centered on a regular staggered grid. The key parameters are the spacing between the bases centers, the standard deviations and the position of the origin basis. In practice, model selection is performed by selecting among different basis settings the one that gives the lowest average distance between the segmentation S and the segmentation obtained by thresholding the prior probability map at the level 0.5.

It has been experimentally assessed that the convergence rate of Alg. 1 is high and even a few iterations give acceptable results. Nevertheless, to guarantee a reasonable computation time, large images are split into overlapping patches where Alg. 1 is performed independently. This approach produces good results since each basis only interacts with very few neighboring bases. The bases from all patches are then combined in the whole image, but bases lying on overlapping regions are weighted with bicubic (resp. tricubic) spline functions for 2D (resp. 3D) images. This approach still leads to a generalized linear model $\sigma(\boldsymbol{\Phi}_n \mathbf{W})$ with $\mathcal{C}^1$ continuity between isoprobability surfaces from neighboring patches.

Once the probabilistic model is fitted, a new segmentation M is generated by thresholding the posterior $p(Z_n | I_n, \mathbf{W})$ at the level 0.5. Two metrics are extracted to measure the adequacy of S with M: the Dice coefficient (DC) $E_D = 2|M \cap S| / \left(|S| + |M|\right)$ and the average asymmetric surface error (ASE) $E_S = d(S, M) = \frac{1}{\partial S} \sum_{x \in \partial S} \min_{y \in \partial M} d(x, y)$ where ∂ denotes the segmentation surface.

We discard the metric $d(M, S)$ as being uninformative since M is not a surrogate ground truth.

4 Results

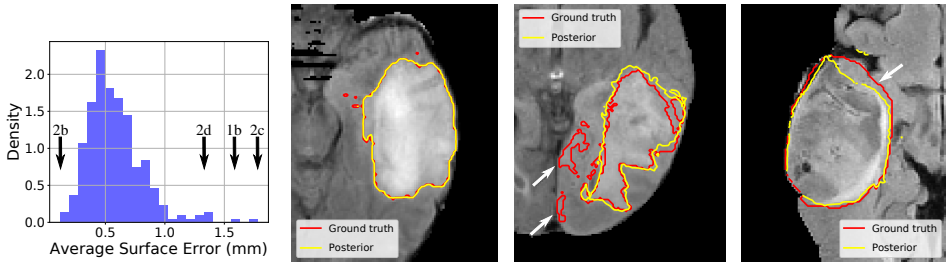
4.1 Quality control of ground truth segmentations

We demonstrate the ability of our algorithm to highlight challenging cases on 285 3D MR segmentations of whole brain tumor from the training set of the BRATS 2017 challenge [6]. The 4 modalities (T1, T1c, T2 and T2 FLAIR) are combined in a multivariate variational mixtures of Student-t distributions with 7 initial components to learn the appearance models of the foreground and background regions defined by the ground truth. Then the posterior is computed using Alg. 1. The distribution of the ASE over the whole dataset (Fig. 2) allows to isolate a dozen of cases at the right tail of the distribution. The case 2c is thus clearly more challenging than the case 2b taken from the left tail. Indeed, the ground truth contour in Fig. 2c is more irregular and could be even questioned because of the very weak intensity variations in some regions (indicated by the arrows). It was maybe extracted through thresholding instead of being manually drawn. This hypothesis is plausible as a thresholding step might have been included in the annotation process [6,3]. Further examples can be seen in the supplementary material. Note that depending on the segmentation task, samples with abnormally low ASE could also be suspicious. Moreover, the average signed distance error gives some indication about the behavior of the segmentation rater. Large negative (resp. positive) average errors probably indicate under (resp. over) segmentations in comparison with M. This is shown in Fig. 1b for under segmentation and 2d for over segmentation. This could be useful to detect rater biases and to improve their delineation performances.

To further enhance the visualisation and interpretability of the results, we can categorize the voxels belonging to the ground truth surface depending on the value of the posterior map at their location and their distance to the isosurface ∂M (Fig. 1c). Explained segmentation are voxels in a neighborhood of 4 mm around ∂M and with a posterior value below 0.9 or above 0.1, for which there is an agreement between the rater and our segmentation model. Mis-aligned segmentations are voxels close to ∂M but with a posterior value above 0.9 or below 0.1 possibly corresponding to a small deviation of the rater around the visible boundary. Finally, voxels that are far from ∂M with a posterior value also above 0.9 or below 0.1 correspond to regions not explained by the probabilistic model and for which a visual review might be worthy.

4.2 Quality control of predicted segmentations

Our algorithm can be of great interest in situations where segmentations are generated by algorithms in the absence of ground truth. For instance, we consider predicted segmentations given by a convolutional neural network (CNN)



(a) (b) (c) (d)

Fig. 2: Study of the BRATS 2017 challenge training set. Distribution of the ASE (2a). Example of a segmentation explained by the model (2b), of a case with regions not explained by the model (2c), and with possible over segmentation (2d), all shown in FLAIR modality.

on 46 test images of the BRATS 2017 challenge as illustrated in Fig. 3. The Dice score computed between the predicted segmentation S and the one obtained by thresholding the posterior map, M, is then compared to the true value obtained by uploading the prediction on the evaluation website of the challenge. Correlations for 3 different tumor compartments are all above 0.69 with few outliers. These results are satisfactory considering that no regression model was learned unlike for example what was proposed in [4].

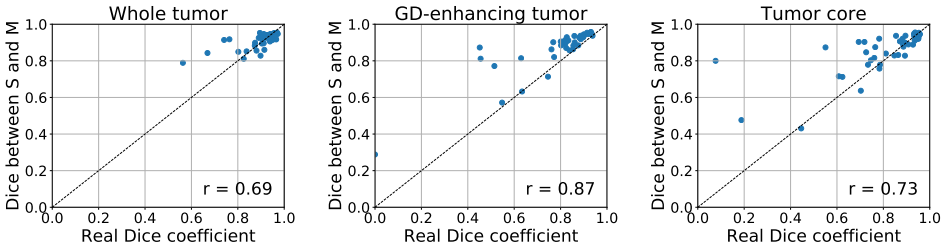
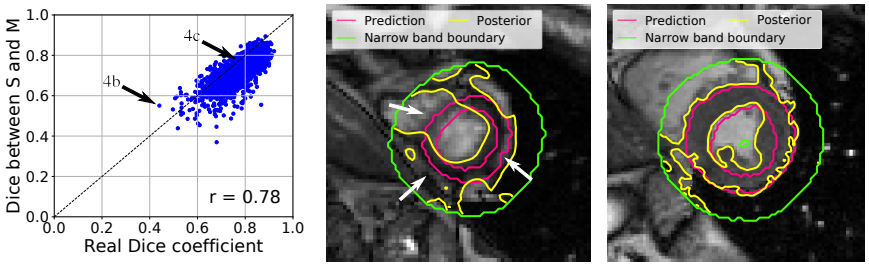


Fig. 3: Real Dice coefficient versus Dice score between the prediction S of the CNN and the probabilistic segmentation M exhibiting good correlation.

We further investigated our algorithm on MR cardiac images from the training set of the ACDC challenge [2]. We evaluate the quality of predicted left-ventricular myocardium segmentations given by a CNN for 100 subjects for which ground truth is available at 2 time points. Each slice is processed individually in 2D due to the large inter-slices distance (1797 slices in total). A good correlation is again observed between the two Dice scores, the first computed

between S and M and the other between S and the ground truth (Fig. 4a). Fig. 4b illustrates a difficult case whereas Fig. 4c shows a well explained case.

Our probabilistic approach is able to automatically distinguish between easy and difficult segmentation cases. Since large segmentation errors are more likely to occur in difficult images rather than easy ones, our unsupervised method is able to provide hints for the cases that are potentially problematic for a segmentation algorithm. Compared to learning-based approach such as [4] or [7] which only output a score, our method provides an explanation of the difficulties through the analysis of the posterior (as highlighted by arrows in Fig. 4b).



(a) (b) (c)

Fig. 4: Study of predicted segmentations by a CNN on the ACDC dataset. Real Dice versus Dice score between the prediction S and the probabilistic segmentation M (4a). Posterior for an ambiguous case (4b) and an easier one (4c).

5 Conclusion

We presented a novel method for quality control assessment of ground truth or predicted segmentations using a Bayesian framework. Our method relies on a generic segmentation model which produces contours of variable smoothness aligned with visible boundaries in the image. Bayesian inference leads to the estimation of all (hyper)parameters without resorting to supervised learning from a subset of data as performed in prior works. Furthermore, the assessment of each segmentation is interpretable. The approach was shown to be a useful tool for quality control assessment for small databases and can indicate the potential performances of segmentations on test data.

Acknowledgements This work was partially funded by the French government, through the UCA^{JEDI} "Investments in the Future" project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01 and supported by the Inria Sophia Antipolis - Méditerranée,"NEF" computation cluster.

References

1. Archambeau, C., Verleysen, M.: Robust Bayesian clustering. Neural Networks **20**(1), 129 – 138 (2007)
2. Bernard, O., Lalande, A., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Trans. on Medical Imaging **37**(11), 2514–2525 (Nov 2018)
3. Jakab, A.: Segmenting Brain Tumors with the Slicer 3D Software. `http://www2.imm.dtu.dk/projects/BRATS2012/Jakab_TumorSegmentation_Manual.pdf` (2012), accessed 16 July 2019
4. Kohlberger, T., Singh, V., et al.: Evaluating segmentation error without ground truth. In: MICCAI 2012. pp. 528–536. Springer Berlin Heidelberg (2012)
5. MacKay, D.J.: Bayesian interpolation. Neural Computation **4**, 415–447 (1991)
6. Menze, B.H., Jakab, A., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. on Medical Imaging **34**(10), 1993–2024 (Oct 2015)
7. Robinson, R., Oktay, O., et al.: Real-time prediction of segmentation quality. In: MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV. pp. 578–585 (2018)
8. Robinson, R., Valindria, V.V., et al.: Automated quality control in image segmentation: application to the UK biobank cardiovascular magnetic resonance imaging study. Journal of Cardiovascular Magnetic Resonance **21**(1), 18 (Mar 2019)
9. Roy, A.G., Conjeti, S., et al.: Bayesian quicknat: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. NeuroImage **195**, 11 – 22 (2019)
10. Valindria, V.V., Lavdas, I., et al.: Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth. IEEE Trans. on Medical Imaging **36**(8), 1597–1606 (Aug 2017)
11. Xu, Y., Berman, D.S., et al.: Automated Quality Control for Segmentation of Myocardial Perfusion SPECT. Journal of Nuclear Medicine **50**(9), 1418–1426 (2009)

— Supplementary material — Unsupervised Quality Control of Image Segmentation based on Bayesian Learning

Benoît Audelan and Hervé Delingette

Université Côte d'Azur, Inria, Epione project-team, Sophia Antipolis, France
benoit.audelan@inria.fr

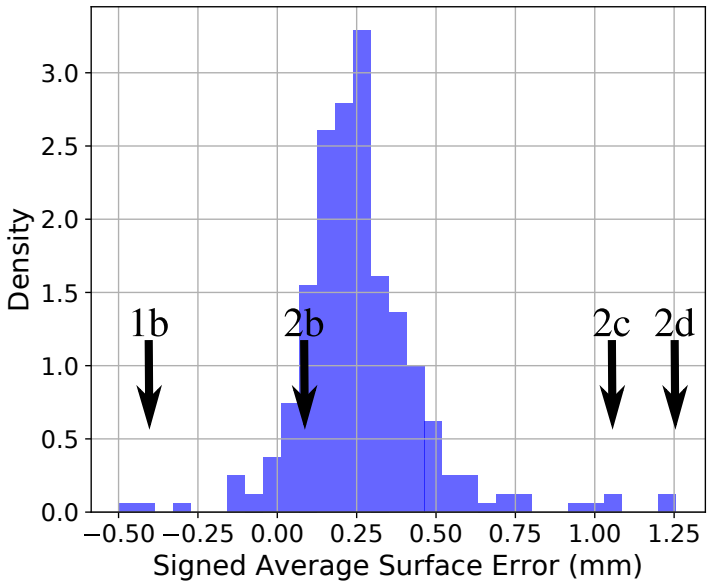
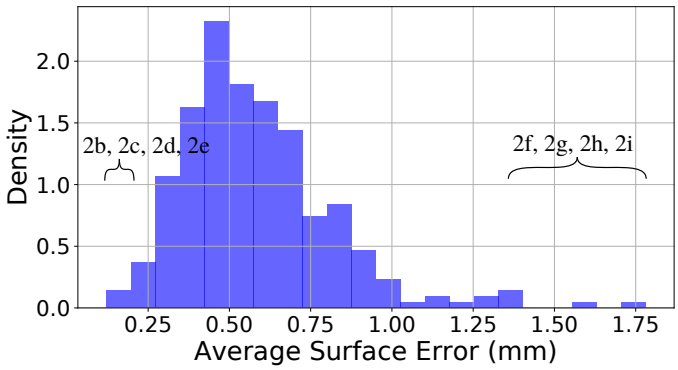
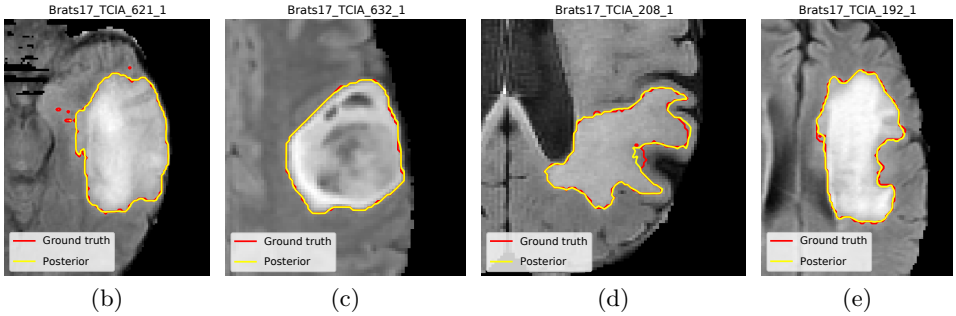


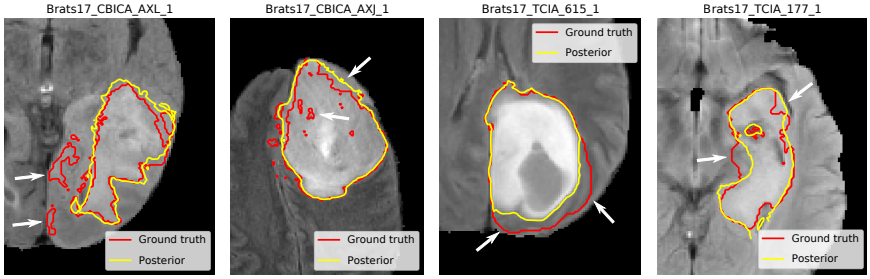
Fig. 1: Distribution of the signed Average Surface Error (ASE) on the training set of the BRATS 2017 challenge, with references to the figures presented in the article. 1b is a case of possible under segmentation whereas 2d could be a case of over segmentation.



(a)



(b) (c) (d) (e)



(f) (g) (h) (i)

Fig. 2: Study of the training set of the BRATS 2017 challenge. Distribution of the Average Surface Error (ASE) (2a). Example of segmentations explained by the model (2b, 2c, 2d and 2e) and of cases with regions not explained by the model highlighted by arrows (2f, 2g, 2h and 2i), all shown in the FLAIR modality.

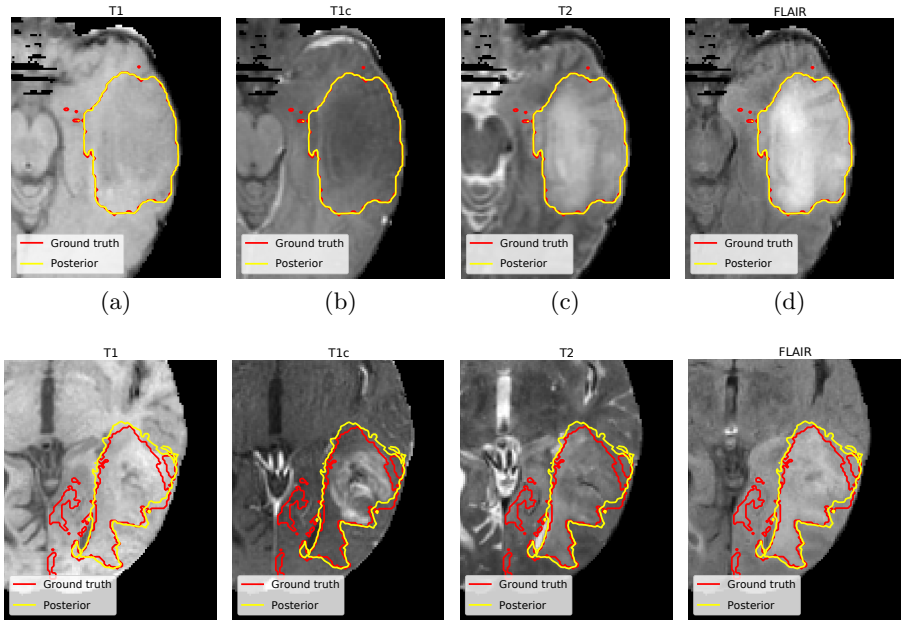


Fig. 3: Posterior and ground truth segmentations shown in the 4 modalities (T1, T1c, T2 and FLAIR) for 2 cases presented in Fig. 2 of the article. The top row is the well explained case 2b while the lower row is the difficult case 2c.